 ISSN 1759-586X

Curating Atmospheric Data for long term use: Infrastructure and Preservation Issues for the Atmospheric Sciences community

SCARP Case Study No. 2

Esther Conway
Digital Curation Centre, Science and Technology facilities Council

DCC SCARP INTERIM CASE STUDY REPORT
Deliverable B4.8.3.2

Version No.	6.21
Status	Final
Date	2 June 2009

Copyright

Catalogue Entry

Title Curating Atmospheric data for long term use
: Infrastructure and Preservation Issues for the Atmospheric Sciences community
Creator Esther Conway (author)
Subject Data curation; Preservation analysis; formats, processes and issues; system development; standards; methodology, and problems overcome; human factors; OAIS; software archiving; representation information
Description Curating
Publisher University of Edinburgh; UKOLN, University of Bath; HATII, University of Glasgow; Science and Technology Facilities Council.
Date 16 March 2009 (creation)
Type Text
Format
Resource Identifier ISSN 1759-586X
Language English
Rights © 2008 DCC, Science and Technology Facilities Council

Citation Guidelines

Conway, E (2009), " Curating Atmospheric data for long term use
: Infrastructure and Preservation Issues for the Atmospheric Sciences community ", SCARP Case Study 2, Digital Curation Centre, Retrieved <date>,from http://www.dcc.ac.uk/scarp

Contents

Executive Summary

DCC SCARP aims to understand disciplinary approaches to data curation by substantial case studies based on an immersive approach. As part of the SCARP project we engaged with a number of archives, including the British Atmospheric Data Centre, the World Data Centre Archive at the Rutherford Appleton Laboratory and the European Incoherent Scatter Scientific Association (EISCAT). We developed a preservation analysis methodology which is discipline independent in application but none the less capable of identifying and drawing out discipline specific preservation requirements and issues. In this case study report we present the methodology along with its application to the Mesospheric Stratospheric Tropospheric (MST) radar dataset, which is currently supported by and accessed through the British Atmospheric Data Centre. We suggest strategies for the long term preservation of the MST data and make recommendations for the wider community.

Study Scope and Contents

The first chapter of the case study gives an overview of the MST data set. We explore the significance of maintaining the long term record and introduce the issue of curation and preservation paying particular attention to the following areas

- Archive history and stewardship
- Nature of data re-use and preservation significance

Chapter Two introduces the preservation analysis methodology

The methodology was developed in response to the challenge of digital preservation. This challenge lies in the need to preserve not only the dataset itself but also the ability it has to deliver knowledge to a future user community. The methodology focuses on the following key areas

- Preliminary investigation of data holdings
- Stakeholder and archive analysis
- Assessment of designated user community
- Identification of Preservation objective
- Creation of preservation information flow diagrams
- Creation of preservation strategies
- Cost/Benefit/Risk analysis

Chapter Three presents the application of the analysis methodology to the MST data set. It considers and recommends preservation strategies which can be adopted by the archive to ensure long term preservation of the data set. Issues explored are

- Good practice in the selection of suitable formats
- Adoption of Climate Forecast conventions and the support of discipline specific vocabularies
- Vulnerable software and the range of preservation strategies available
- Archiving of web based resources
- Grey material, the persistence of repositories and the provision of good quality long term references
- Acceptable risks and persistent skill set of a user community.

We then consider and recommend preservation strategies to the archive for identified preservation risks.

Chapter Four explores the implications of issues raised by this case study outside the immediate needs of the MST data set, and makes recommendation to the DCC in order to support curation and preservation outside the atmospheric science discipline in order to serve the wider community.

Report Conclusions and recommendations

This is a report from the DCC SCARP project; the opinions and recommendations herein are those of the author, and do not represent the positions of STFC, The University of bath or the University of Edinburgh (the DCC SCARP partner institutions). The report's recommendations will be considered by the DCC and appropriate actions taken following discussion of strategy and resource implication. The conclusions and recommendations for the Archive, DCC and wider community are listed below.

Recommendations for consideration by the Archive

It is our recommendation that: the Archive should create a preservation plan based on a cost benefit and risk assessment of the available strategies. Publish this along with an assessment of both preservation objectives and the designated community, for public scrutiny and comment. Review this plan periodically, adjusting it in response to environmental changes and improvements in preservation techniques. Carry out necessary preservation actions, and create a "logical Archival Information Package" during this current period of quality management and active use while the resources and information are still obtainable. This should allow the data to be retired from active management, or transferred to another organisation with greatly reduced preservation risk.

Recommendations for consideration by the DCC and the wider community

There is a need to support preservation analysis and planning at the data set level and establish a process which is comprehensive and aware of all elements required for the re-use of data in the long term. We also identified areas where archives may benefit from external support in order carry out appropriate analysis, strategy selection and preservation action.

Recommendation ~1 Wider application, trialling and further development of the preservation analysis methodology outlined here would be desirable to test its validity in a broader range of disciplines and organisational settings. In addition the production of training materials and support for archivists who wish to adopt our approach for data preservation would be of benefit.

Recommendation ~ 2 We view the preservation analysis methodology as complimentary to repository planning, audit and certification activities. Further investigation is needed into how the results of preservation analysis could be fed into audit and risk analysis assessments such as Drambora. Integration of preservation analysis with other digital preservation practices is necessary to provide archives caring for scientific data sets with the full arsenal of tools and techniques necessary to rise to the challenge of digital preservation.

Recommendation ~3 Archives can find it difficult to articulate and specify reasons for the preservation of data. We recommend that the DCC develops further guidance on setting preservation objectives and establishing valid business cases for the preservation of scientific data.

Recommendation ~4 Archives need to establish the skill and knowledge base they should monitor in their "designated community", in order to ensure data re-use. The DCC should investigate this area further, and provide guidance and assessment tools to facilitate the meaningful definition and monitoring of such a designated community.

Recommendation ~ 5 Persistent access to grey literature that supports data re-use is an important issue. Advice on approaches for the deposit and citation of such material would be a valuable service for archives.

Recommendation ~ 5a Similarly, the data curation community would benefit from an notification service for repositories that are in danger of closing or whose content is being migrated to another repository to ensure persistent access to required content.

Recommendation ~6 DCC could offer an advisory service which recommends or provides information on preservation strategies available to archives. It could additionally provide quality assurance and testing for representation information deposited in the DCC Registry/Repository of Representation Information (RRORI).

Recommendation ~7 The DCC or another identified organisation could provide an archiving service and/or assistance for web based collections of representation information and preservation description information.

Recommendation ~8 The MST data has benefitted from many good data management practices recommended through the British Atmospheric Data Centre. Other data sets from

outside the atmospheric sciences could benefit from similar approaches. Self-describing, well documented data formats such as NetCDF, semantic control through CF standard name conventions, and software development initiatives are just some examples of practices which could be transferred outside the discipline. The DCC should play an instrumental role in transferring good practices between disciplines.

1 The Mesospheric Stratospheric Tropospheric Data Set

Fig. 1.1 MST radar site at Aberystwyth

1.1 Introduction and overview of the MST data set

The Mesosphere-Stratosphere-Troposphere (MST) Radar at Aberystwyth is the UK's most powerful and versatile wind-profiling instrument. It is unique in being able to provide continuous measurements of the three-dimensional wind vector over the altitude range 2-20 km at high resolution (typically 300m in altitude and a few minutes in time). It can also provide information about atmospheric stability, turbulence, humidity and rainfall. It is therefore ideally suited for studying everything from small-scale atmospheric phenomena through to large-scale weather systems. Wind-profile data are supplied to the Met Office, [1] for numerical weather prediction purposes, through a commercial contract. Upper air input from the Aberystwyth area has been found to have a significant impact on improving longer range forecasts.

The mission of the Facility is:

- To operate the radar on behalf of the UK atmospheric science community
- To operate, and host, instruments whose observations complement those made by the MST radar
- To facilitate the analysis and interpretation of the data

Fig. 1.2 location of MST radar site

It is a 46.5 MHz pulsed Doppler radar ideally suited for studies of atmospheric winds, waves and turbulence. It is run predominantly in the ST mode (approximately 2–20 km altitude) for which MST radars are unique in their ability to give continuous measurements of the three dimensional wind vector at high resolution (typically 2–3 minutes in time and 300m in altitude).

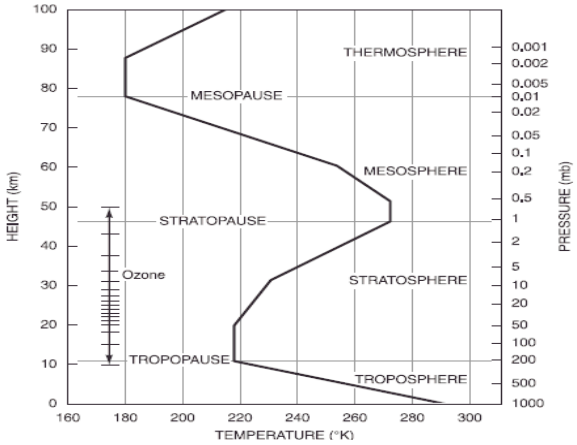


Fig. 1.3 Structure of the Atmosphere

Wind-profiling radar systems can be designed to operate at frequencies anywhere between 40 and 1400 MHz In practice, however, they are restricted to frequencies around 50, 400 and 1000 MHz

Fig. 1.4 Instruments at MST site

Doppler Beam Swinging (DBS) involves making observations in a cyclic sequence of vertical and near-vertical beam pointing directions. The 'targets', from which small fractions of the pulsed radar signals are returned, are irregularities of atmospheric refractive index, which cause back-scattering (so-called 'clear-air' returns), and hydrometeors, which give rise to Raleigh scattering. The scattered signal is Doppler-shifted according to the radial component of the target's velocity i.e. that along the radar beams pointing direction. Profiling is achieved by sampling the radar return signals as a function of delay from the time of the transmitted pulse; the transmitted pulse length determines the range resolution.

Wind profiler radar returns are parameterised by their signal powers and spectral widths (i.e. the variance of scattered velocities about the mean) in addition to their Doppler shifts. This information can be used, under certain circumstances, to provide additional information about the atmospheric static stability (thus allowing monitoring of the altitude and sharpness of the tropopause), humidity fields and turbulence (of at least moderate intensity).

MST radar ST –Mode Wind Quick Look Plot

Shows vertical shear of horizontal wind, beam-broadening-corrected (vertical beam) spectral width, aspect sensitivity, and tropopause altitude.

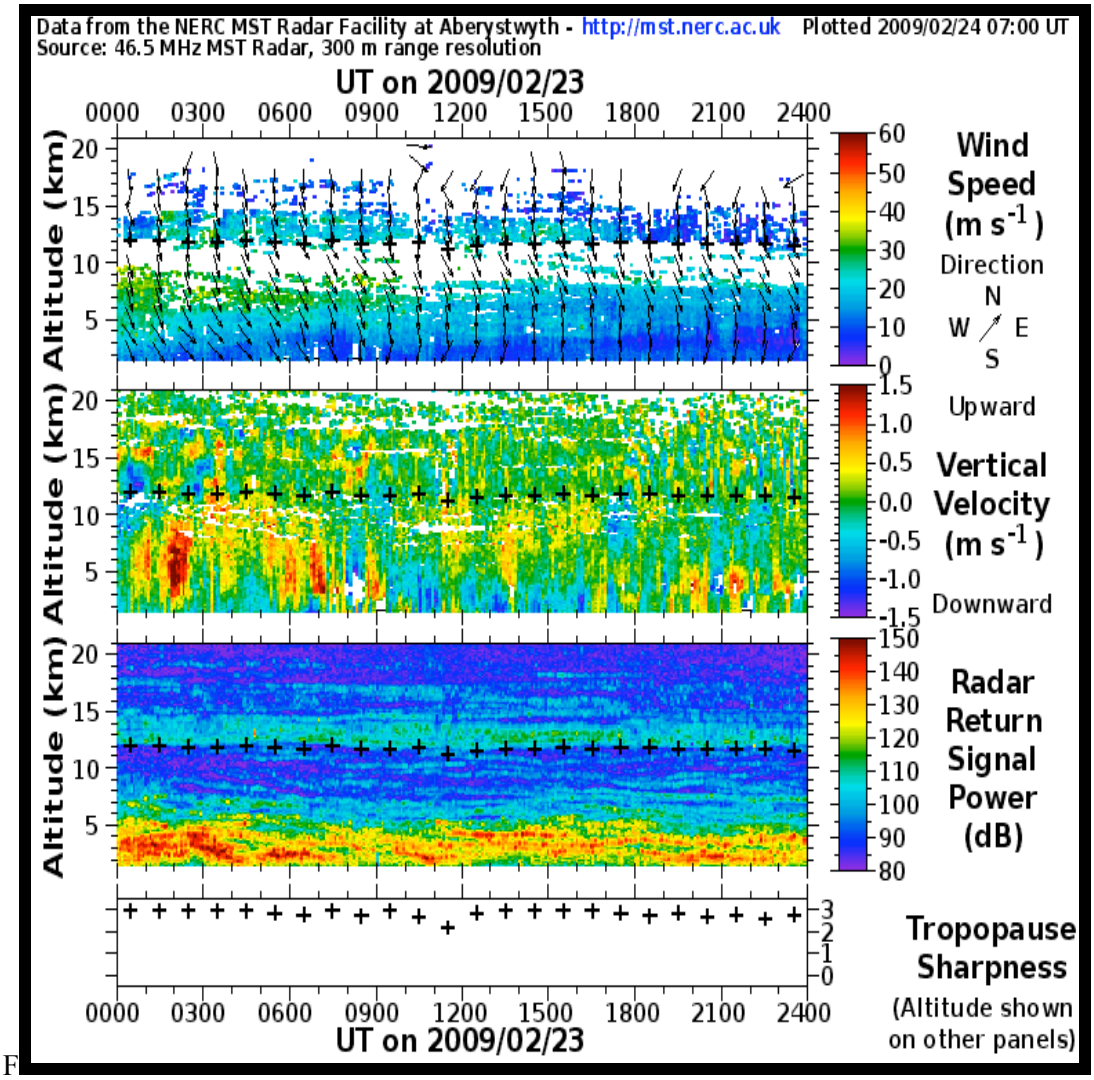


Fig. 1.5 MST radar ST –Mode Wind Quick Look Plot

MST radar ST –Mode Turbulence Quick Look Plot

Shows vertical shear of horizontal wind, beam-broadening-corrected (vertical beam) spectral width, aspect sensitivity, and tropopause altitude.

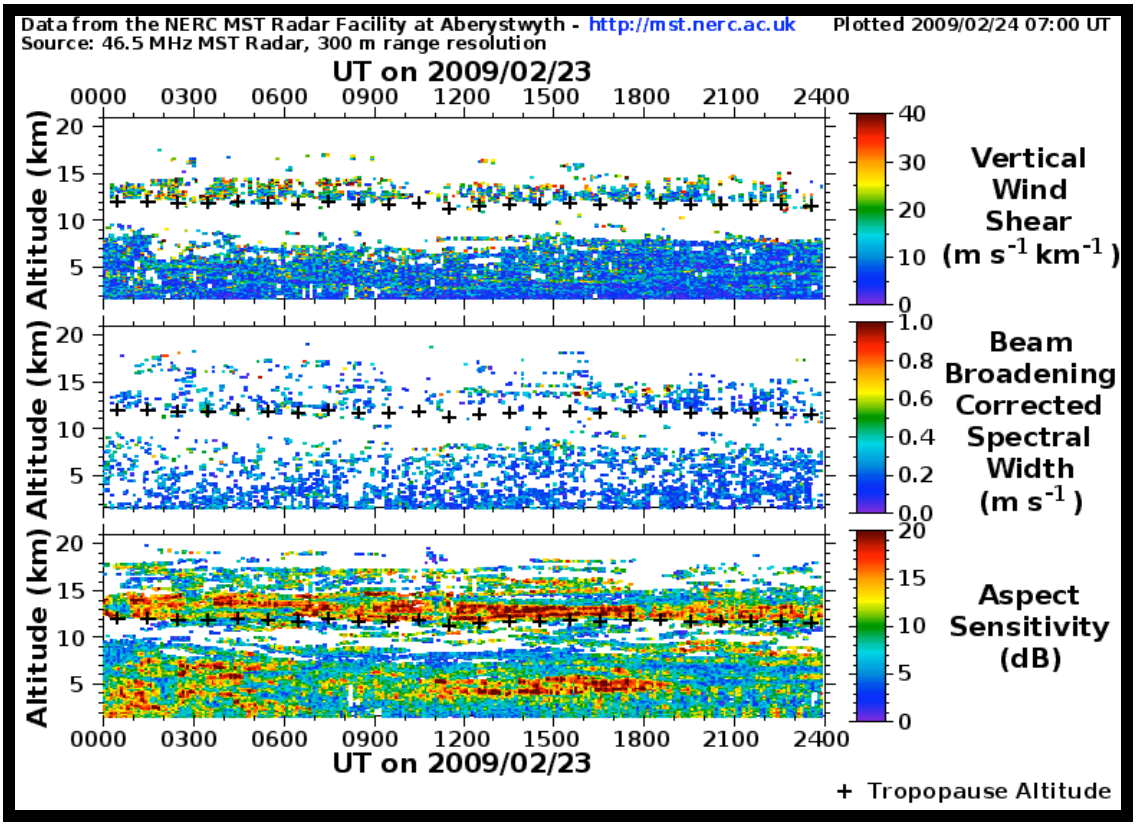


Fig. 1.6 MST radar ST –Mode Turbulence Quick Look Plot

MST Radar - ST mode – diagnostics

Shows where a secondary radial chain exists (i.e. evidence of structure in "unwanted" signal components), the complementary-beam horizontal-velocity continuity factor, the theta_s compensation factor (which has been applied to the wind speed to compensate for the effects of aspect sensitivity), and the tropopause altitude. These plots are intended for diagnostics purposes only

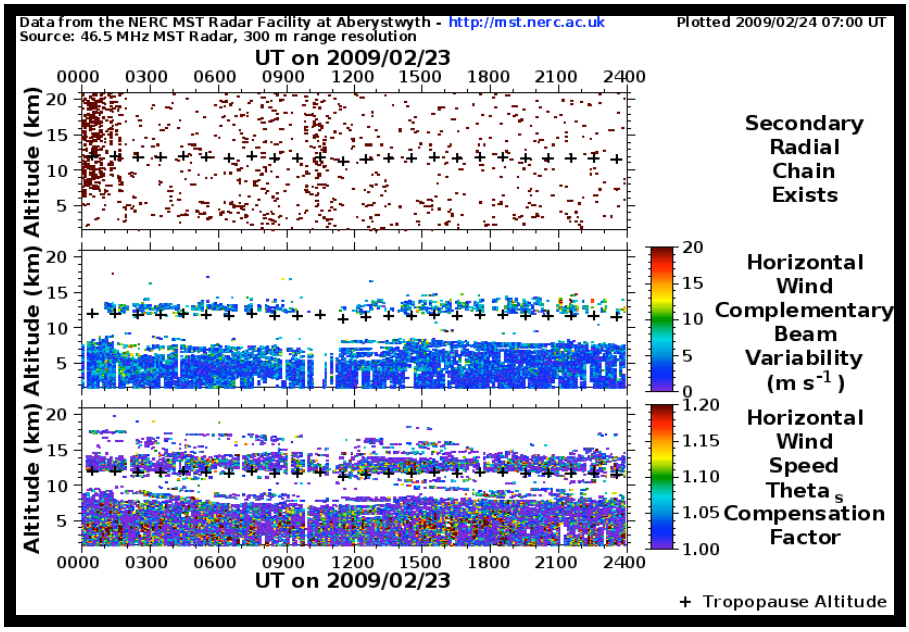


Fig. 1.7 MST Radar - ST mode – diagnostics

MST Radar - M mode

Shows the (vertical beam) radar return signal

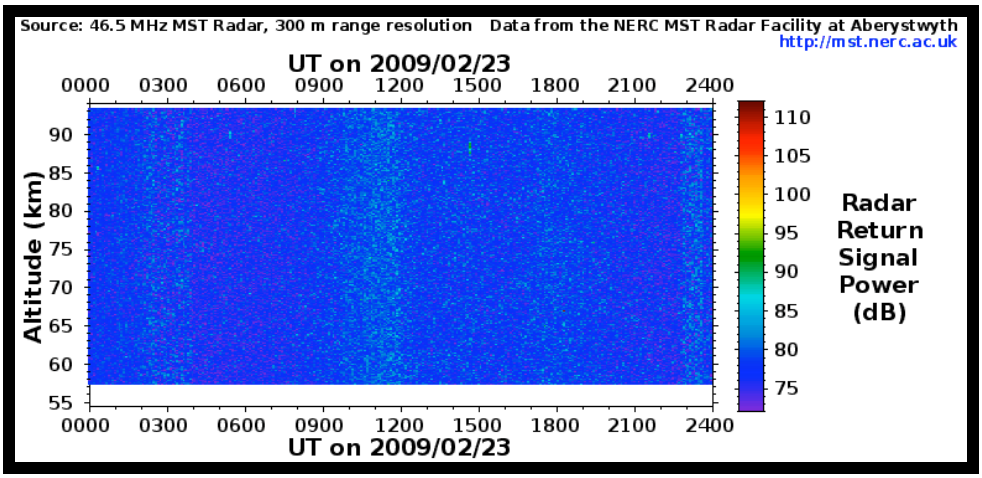


Fig. 1.8 MST Radar – M mode

1.2 Archive history and stewardship of MST Data

The MST data are produced by a facility funded by the Natural Environment Research Council [2]. It is managed by the Space Science and Technology department at the Rutherford Appleton Laboratory [3] in collaboration with The Met Office, The School of Earth Atmospheric Science at the University of Manchester and Aberystwyth University.

Data can currently accessed via the British Atmospheric Data Centre[4]. The BADC is one of the centres and facilities in the NERC Centres for Atmospheric Sciences, NCAS. NCAS [5] carries out the core research programme in atmospheric science funded by NERC. The BADC is one of seven designated data centres established to carry out the NERC data policy. This policy is outlined in the NERC Data Policy Handbook [6], which outlines the responsibilities of both NERC funded researchers and the NERC designated data centres.

The BADC has substantial data holdings of its own and also provides information and links to data held by other data centres. The data held at the BADC are of two types:

1. Datasets produced by NERC-funded projects; these datasets are of high priority since the BADC may be the only long-term archive of the data.
2. Third party datasets that are required by a large section of the UK atmospheric research community and are most efficiently made available through one location

The MST data would be considered to be of the first type.

The BADC aims to be up to date with both the technology of data management and the science of the community it serves. As a consequence it has influence on data management practices such as adoption of data formats and standards such the climate forecast conventions which we will discuss later on in this case study.

The MST radar data set is extremely well documented and tightly managed. Access to the data is restricted, with end users required to report back on how they have used the data. The Archivist is the key manager of these data for a number of reasons

- He is also the project scientist involved in production of the data
- He is a field expert and practising scientist in close contact with relevant scientific organisations, publishing at and attending conferences.
- He additionally provides support, runs and keeps records of user group meetings.
- He provides reporting to the funding bodies.

The result has been not only the opportunity to preserve atmospheric measurements made, but also the current knowledge surrounding atmospheric behaviour evident within the data. We will revisit this in section 3 when we suggest the preservation objectives for the Archive. Listed in section 1.4 are some examples of the atmospheric phenomena which can be observed through MST data; a bibliography of resulting journal articles can be found at the MST website [7] http://mst.nerc.ac.uk/publications_by_year.html.

1.3 Nature of data use and preservation significance

The MST data can be used in a number of ways. Simple extractions of the following parameters will provide a snapshot of a particular type of atmospheric behaviour

- Wind Speed and Direction
- Vertical Velocity
- Radar Return Signal Power
- Tropopause Sharpness
- Tropopause Altitude
- Vertical Wind Sheer
- Beam Broadening Corrected Spectral Width
- Aspect Sensitivity
- Secondary Radial Chain
- Horizontal Wind Complementary Beam Variability
- Horizontal Wind Speed Thetas Compensation Factor

However much of the meaningful use requires much more than this simple extraction and viewing of parameters. Use of atmospheric data may involve format conversion to make it interoperable with other data sets. Specialist visualisation of time related data may be needed to study some kinds of atmospheric behaviour. Interpolation, subsetting or different forms of statistical analysis may also be employed. Identification of atmospheric phenomena and behaviours is achieved through the creation or application of data to established models of dynamic atmospheric systems. The same data may be used to study different aspects of atmospheric behaviour; listed below are some which we identified in peer reviewed literature resulting from studies which employed the MST data.

Precipitation

Clouds contain moisture; when the droplets in clouds coalesce they become sufficiently large to cause precipitation. A recent evolution of knowledge surrounding the MST radar data allows the data to be used to study precipitation

Convection

Convection is the transfer of heat by movement within a substance. The MST radar data permits you to study the convective circulation of air within the atmosphere.

Gravity Waves

Gravity waves are generated in the troposphere by frontal systems or by airflow over mountains. The geographic position of the MST radar site is ideal for studying this phenomenon.

Fig.1.9 Cloud formation due to gravity wave of the coast of Aberystwyth

Rossby Waves

Rossby waves are a subset of inertial waves. These atmospheric waves are large scale motions with wavelengths of up to 6000 km. The continual monitoring of a discrete region of the atmosphere over a long period allows for the analysis of such waves.

Mesoscale and Microscale Structures

The frequency of observation and the resolution of the MST radar also permit analysis to be carried on mesoscale (~50km) and microscale (atmospheric lasting a matter of minutes) structures.

Fallstreak Clouds

Cloud formations can be associated with atmospheric conditions such as turbulence and waves which the MST radar is capable of observing.

Ozone Layering

Atmospheric dynamics can also be correlated with chemical composition of the atmosphere

Preservation significance of the data set

The importance of the MST dataset lies in the fact that it is an irreplaceable earth observational record: once lost, these data cannot be replaced by the repetition of the experiment. The dataset is valuable because it

- Contains data from the UK's most powerful and versatile wind-profiling instrument
- Provides information about atmospheric stability, turbulence, humidity fields, precipitation and variety of atmospheric phenomena
- Contains measurements of winds up to many kilometres from the ground
- Contains a record of winds sampled continuously, with a cycle time of a few minutes over a long period of time
- Provides a record not only of the horizontal but also the vertical air velocity which additionally has high temporal and spatial resolution.

The challenge of digital preservation lies in the need to preserve not only the dataset itself but also the ability it has to deliver knowledge to a future user community. In order to carry out meaningful preservation we need to ensure that future users are equipped with the necessary information to re-use the data. We will spend the rest of this case study applying an analysis methodology and examining the preservation issues and solutions available for this data set.

2 Preservation Analysis Methodology

In this case study we sought to incorporate a number of analysis techniques tools and methods into an overall process capable of producing an actionable preservation plan for scientific data archives. The workflow below illustrates the stages of this analysis methodology

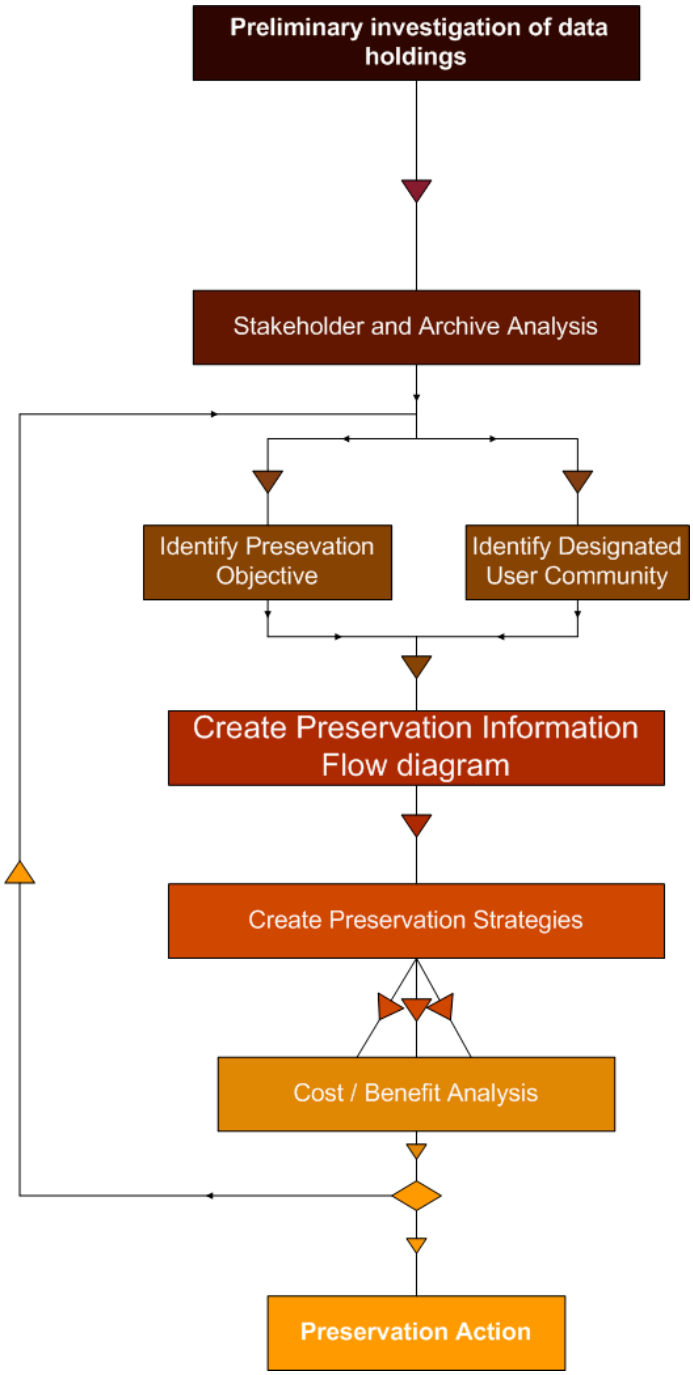


Fig. 2.1 Preservation analysis workflow DCC SCARP project 2008

2.1 Preliminary investigation of data holdings

We propose to use the CASPAR questionnaire [8], which contains keys questions which allow us to carry out a preliminary investigation into an archive's data holdings. The CASPAR questionnaire is strongly guided by OAIS [9] and the CASPAR architecture. It lays out 14 key questions which critically allow us to:

- Understand the information extracted by users from data
- Identify Preservation Description and Representation Information
- Develop a clearer understanding of the data and what is necessary for effective re-use
- Understand relationships between the data files and what constitutes a digital object within the archive.

While it is appreciated that this questionnaire is not an exhaustive list of questions which one may need to ask about a preservation target it still provides sufficient information to commence the analysis process

The Full questionnaire can be found on the CASPAR website at http://www.casparpreserves.eu/Members/metaware/ReferenceDocuments/caspar-questionnaire and full results form the Questionnaire from the Ionosonde WDC holdings[10] can be found at www.**caspar**preserves.eu/other-**caspar**-products/other-**caspar**-products/**ionosonde-case-study**.pdf

2.2 Archive Stakeholder, Evolution and Management

Stakeholder Analysis

After carrying out the questionnaire process for each data archive it became necessary to carry out a stakeholder analysis for these archives. This is due to

- Stakeholders having differing views of the knowledge a data set was capable of providing an end user,
- Stakeholders identifying different end users who possess varying skill sets and knowledge base,
- Stakeholders producing or being custodians of different information vital for re-use of the data.

Stakeholder Categories

After inspecting a number of datasets the following categories of stakeholder were felt to be most appropriate for science data.

Funding Bodies

Every digital archive will have some form of funding body associated with it in order to provide the resources to collect and maintain the data. During its lifetime, the funding for a data set may be received from several bodies generating rich documentation which explains the scientific purpose of the dataset, and how data use has evolved over time. These documents can take the form of experimental proposals which will explain the original intent of the experiment/observation, institutional reports which state the intent of maintaining supply of the data to a scientific community, and reports which show successful scientific output. It is worth noting the limits of such documentation, as it will omit scientific use outside the remit of the organisation. In the case of the science archives we engaged with, we observed how different Research Councils are interested in different regions of the atmosphere resulting in the documentation of some areas of scientific investigation not being included in reports.

Scientific Organisations

Scientific organisations such as university departments, national or international institutes and laboratories, are frequently associated with datasets. They tend to work within a particular branch of science and can provide a great deal of detailed information on how a dataset can support that particular area of scientific investigation, providing for example software support materials and field-specific bibliographies. However, these scientific organisations, whilst being an excellent source for support for that area of scientific discovery, will naturally neglect other disciplines. In data archives this was particularly evident for emerging and specialist areas of scientific investigation, where much knowledge was still embedded in the data-using scientific groups, and there tended not to be mature documentation supporting the data.

Data Producers

Every dataset will have an individual scientist, or group of scientists responsible for its production. In addition to the scientific intent recorded in an experimental proposal, they will also hold other information and make additional observations at the time of the experiment/observation which can enhance use of the data. These could be event associations with other phenomena, for example lighting strikes and ionisation of a region of the atmosphere, or identification of recurrent patterns which merit further investigation. In the case of the MST dataset we see how the project scientist discovered that a signal signature due to precipitation was present in a dataset traditionally used for wind profiling. In this instance the scientist was able to study this and publish his finding in his paper on "VHF signal power suppression in stratiform and convective precipitation" [11]. This type of material has a tendency not to be formally recorded, sometimes manifesting itself in wiki and web logs. There is concern that much of this type of information is at high risk of loss.

Scientists in the Community

This collection of scientists is the most diverse and distributed. Indeed other groups may be considered to be a subgroup of scientists as their opinion will have been ultimately informed by the larger scientific community. Except in the circumstance of highly specialised datasets with discrete user communities, we would also expect a full survey of the wider data users to be completely unrealistic. The ability to capture such information from an active data-using community would be greatly enhanced by the developed of annotation systems such as AstroDAS [12] which permit the annotation of astronomical data allowing for scientific assertions to be captured. Projects such as CLADDIER [13] and OJIMS [14] are also developing ways of referencing and kite-marking datasets which will potentially ease the discovery of knowledge associated with datasets.

Data Archivist

The Archivist is the group or individual who is the current custodian of the data. The extent to which they have interacted with other stakeholder groups and extracted knowledge requirement with its associated information will be highly dependent on the resources available to, the motivations, background and personal bias of the individual Archivist

Archive Evolution and Management

In addition to familiarizing oneself with the stakeholders from the different categories it was additionally beneficial to understand how an archive has evolved and been managed. This can used to illuminate the different uses of data over time and the production of associated representation information vital for that type of use

The diagram below is a graphical representation of the awareness the different stakeholders have of data use by scientists and their relationships to each other.

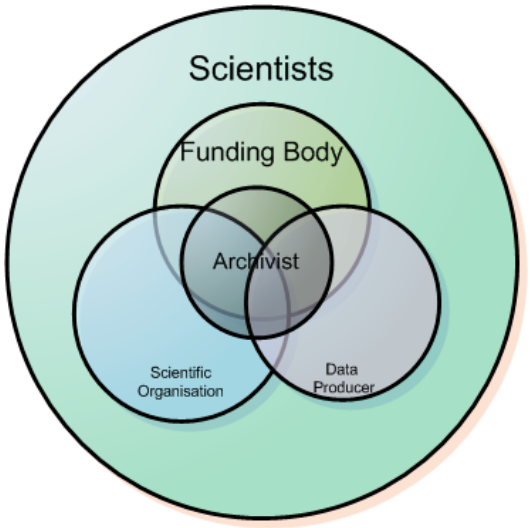


Fig. 2.2 Key Science data stakeholders

We looked at the following sorts of factors which influenced the use and re-use of data over time

- Birth and development of a science
- Events which influence data use such as the Second World War, or global warming
- Development of countries' technologies, and the emergence of global networks
- Publication of journals, technical manuals, interpretative handbooks, conference proceedings, minutes of user group meetings, software etc.
- Emergence of branches of science and associated organisations
- Stewardship of data and the influence of different custodians.

This is not an exhaustive list as many factors influencing data re-use are domain specific, as also is the categorization of the stakeholders. The generic principle of carrying out stakeholder characterization and the identification of factors will be domain independent. Naturally most of these can only be expected to be dealt with in the most cursory way in any practical study nevertheless even this can be extremely important in understanding the situation.

2.3 Defining a preservation objective

The analysis carried out before this point may present one with a natural, easily defined preservation objective, or alternatively there may be a greater number of options which overlap and are more difficult to define. It is important to note that this type of analysis cannot advise which preservation option to choose but merely clarifies the options available.

Preservation objectives should be

- Specific: well defined and clear to anyone with a basic knowledge of the domain
- Actionable: the objective should be currently achievable. It is important to note that information ultimately to be extracted by a future user cannot be predicted and therefore we should not attempt to "predict the future"
- Measurable: it is critical to be able to know when the objective has been attained in order to assess if any preservation strategy developed is adequate.

2.4 Defining a designated user community

The Designated Community is defined in OAIS as **"**An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities"

An archive must define the Designated Community for which it is preserves some digitally encoded information, and must therefore create AIP's with Representation Information appropriate for that Designated Community.

The Designated Community will possess a skills and knowledge base which allows its members to successfully interact with a set of information which has been stored with an AIP, in order to extract required knowledge or recreate the required performance or behaviour. In common with the preservation objective, the analysis up to this point may suggest a range of community groups the archive will serve.

The definition of the skill set is vital as it determines the limit to the amount of information which must be stored with an AIP in order to satisfy a preservation objective. In order to do this the definition of the Designated Community must be

- Clear with sufficient detail to permit meaningful decisions to made regarding information requirements for effective re-use of the data.
- Realistic and stable in so far as there is reasonable confidence in the persistence of the knowledge base and skill set.

While the need to define the Designated Community of users is universal, the nature of a knowledge and skill set will tend to be domain specific. The following are typical examples from atmospheric science

- Ability of a community to successfully operate software i.e. knowledge of correct syntax to input commands into a UNIX command line.
- Ability to utilise appropriate analysis techniques with data to remove background noise or identify specific phenomena
- Comprehension of community vocabularies
- Appreciation of different scientific techniques employed during the production of data, their limitations and comparative success rates for picking up desired phenomena
- Knowledge of atmospheric events or processes which may be affecting the atmospheric state being measured within a data set.

It is the appraisal of this knowledge skills base as a permanent attribute of the designated user community which will determine whether it is necessary to preserve such information by including it with an AIP (Archival Information Package).

2.5 Preservation information flows

The OAIS Reference Model specifies that within an archival system, a data item has a number of different information items associated with it, each performing a different role in the preservation process. OAIS asserts that the preservation objective for a designated community is satisfied when each component of the OAIS reference model has been adequately populated with sufficient information (ie, this will be sufficient for effective re-use). We examine these information types in turn.

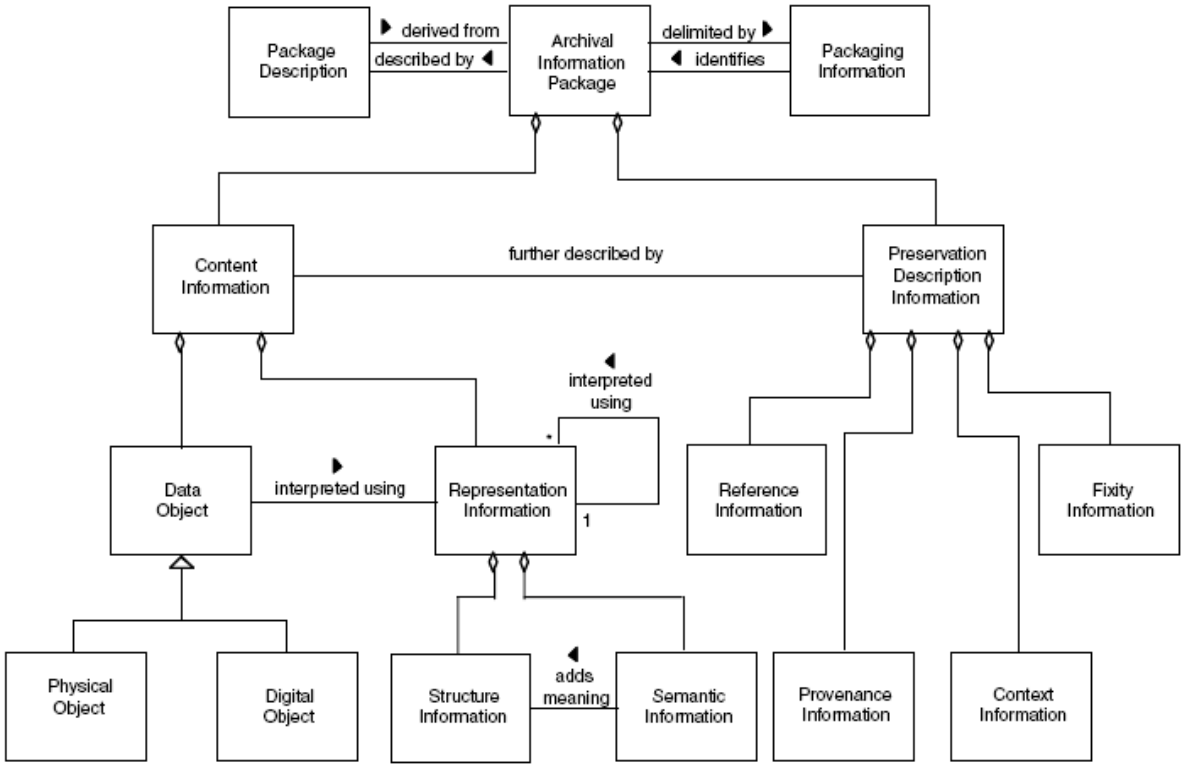


Fig. 2.3 – OAIS (Reference Model for an Open Archival Information System) Information Model

Preservation Description Information

OAIS specifies that information be provided to describe the data set with properties required for preservation. Such Preservation Description Information comes in four types.

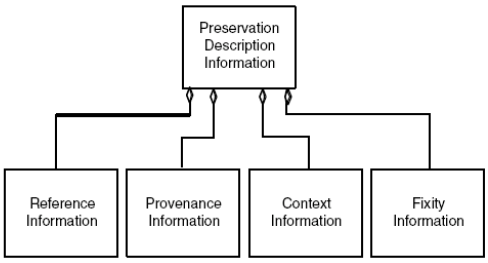


Fig. 2.4 – OAIS (Reference Model for an Open Archival Information System)

Preservation Description Information

Reference Information: Reference information assigns identifiers within identifier schemes to the data, and is independent of the preservation objective.

Context: Context describes the relationship between data and its environment. The most basic knowledge requirement would need very basic context information, where for example a snapshot of the atmosphere at a particular time and location is required. However when this knowledge requirement expands, eg for example tracking large scale atmospheric phenomena, one may need to establish temporal and spatial relationships between files in a dataset in order to do this. For example a funding institution such as the Met Office requires this type of snapshot information to feed into predictive models, but an individual scientist studying Mountain waves would wish to study a type of phenomena which takes time to pass over the MST radar site and which are approximately 8 miles in length. This may expand again if the information in the data set is required to interact with external data sets, which may for example mean mapping to a different co-ordinate system, such as heliocentric and geocentric systems in astronomical data, to allow for such interoperability.

Provenance: Provenance information documents the history of the data, what actions were performed on the data, by whom and when. Provenance may for simple requirements be viewed as a special type of context information, as in the case of snapshot type data for the Met office.

However more detailed provenance information may be required if factors such manual or automated scaling may affect the data quality. It was noted that the appearance of a particular phenomena such as the occurrence of a sporadic E-Layer, was most reliably identified by a skilled manual scaler visually inspecting ionograms. This means that future scientists conducting research related to the appearance of the sporadic E-Layer should use only data which had been through this type of process. Physical factors such as the type of instrument and its use can have an effect, eg in the EISCAT [14] archive we see how the scientific objective of special programme experiments influence the instrument's mode of operation, and in some cases the EISCAT instrument has been operated to respond to events of geophysical interest such as proton events or earthward directed coronal mass ejections. If the preservation scope of an archive encompasses investigation of such events or scientific objectives, adequate provenance is necessary for the discovery and use of the data.

The data may additionally need to be from trusted institutions to ensure a desired level of authenticity. Currently the ingest of data into the Rutherford Appleton Laboratory WDC [15] archive is highly reliant on the Archivist's appraisal of trusted producers. Maintenance of such an archive needs the addition of provenance information relating to these producers to demonstrate the required authenticity to future users.

Fixity: Fixity information documents the authenticity mechanisms for the data within the archive. Fixity information may generally be considered to be independent of knowledge requirements except in the case where a specified level of authenticity is part of that requirement. For example if atmospheric data were required as evidence of a pollution event in a legal case, it may be necessary to demonstrate the data had not been tampered with, by means such as a digital signature.

Representation Information

Representation information in OAIS describes how the information is signified by the data, including what semantic content is being represented and how that is physically rendered in the data format. Representation Information has three main types.

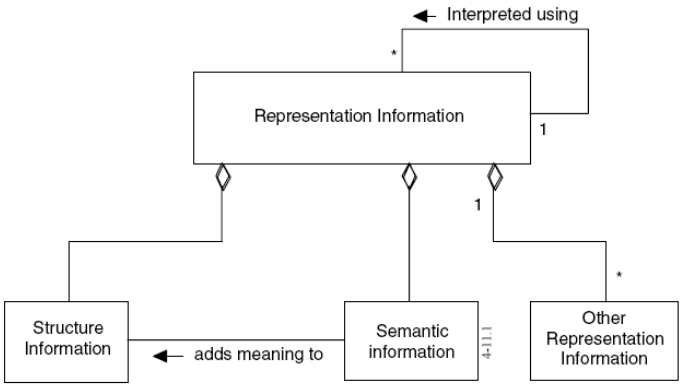


Fig. 2.5 – OAIS (Reference Model for an Open Archival Information System)

Representation Information Model

Structure Information: The required structure information is the minimum information needed to extract and correctly identify the required parameters. The knowledge to be extracted from the data set determines the required level of processed data and the parameters from that level of data. These parameters may form part of the content or indeed part of the provenance information, for example in the case of station identification or mode of operation for an instrument.

Semantic Information: The level of semantic information required varies according to the level of understanding, interpretation and authenticity which is needed to be attached to the extracted parameter. This ranges from simple definition from communities such as in the ionospheric science use case, CF naming conventions [16], URSII parameter definitions [17] to extensive documents such as the URSII handbook of Ionogram interpretations [18]. These semantic definitions may additionally evolve over time as user community vocabularies shift.

Other Representation Information, including Higher Level Knowledge:

The amount of additional "other" materials needed tends to be the most explosive in reaction to the expansion of the preservation objective. Typical examples include

- Software including scientific models
- Code documentation, description of algorithms
- Support materials for operation of software
- Web Pages including support materials, educational materials, non technical documents for consumption by a general audience, information packs and background documents
- Subject specific bibliographies and texts.

OAIS Preservation information flow diagrams

An OAIS preservation information flow diagram is a graphical representation and analysis tool, which is a hybrid of an information flow diagram and the OAIS reference model. We developed this technique on the SCARP project as a way of logically identifying the information which needs to be preserved to satisfy a preservation objective for a specified Designated Community. It gives the addition benefit of providing a convenient format to facilitate group discussion over preservation plans and strategies.

Elements of OAIS Preservation information flow diagrams

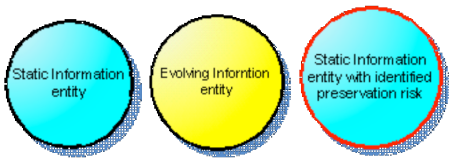
Standard OAIS information model components of an AIP. These are the standard components of an AIP as described above. All information entities must be mapped to at least one of the following components within an AIP.

Information Objects

An information object is a physical unit of information suitable for deposit within an AIP as it currently exists. An information object must have the following attributes

- Name
- Description of the information contained by the entity which is vital for the preservation objective e.g. a piece of software contains structural information and algorithms for the processing of data within its code
- Description of the format i.e. website, PDF, database or software
- Assessment of preservation risks and dependencies.

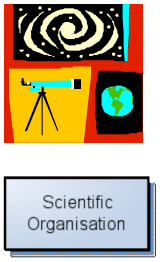
Notation used



Stakeholder entities

A stakeholder entity is the named custodian of the required Information entity.

Notation used

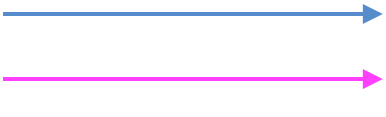


Supply Relationship

The supply relationship should simply be an indicator of any impediment to the current supply of an information entity such as an embargo or assertion of copyright. The attributes of a supply relationship are

- Supply possible (Yes/No)
- Description of supply impediment.

Notation used



Supply possible: Yes

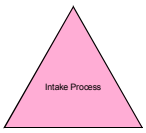
Supply possible: No

***Supply* Process**

The supply process is any process carried out on information supplied by the stakeholder in order to produce the information object. Its attributes are

- Name
- Description of process e.g. dump of a database table into a csv file, archiving of public website or reformatting of data files.

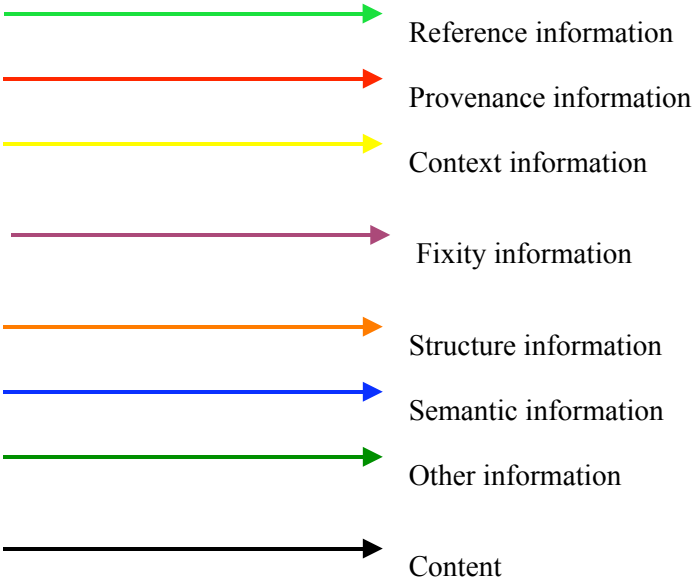
Notation used



Packaging relationship

The only required attribute of the packaging relationship is that it links an Information entity to at least one standard OAIS Reference Model component of an AIP. However many implementations of packaging such as XFDU[16] require additional information.

Notation used



Information object dependency relationships

The information object dependency relationship connects two information objects. If preservation action is carried out on one object there is an impact on another object with a dependency. For example if a piece of software is identified to be at preservation risk and deconstructed to a structural format and analysis algorithm descriptions, the software user manual will be flagged up by the dependency relationship and may be removed on the basis that this information is now irrelevant.

Notation used

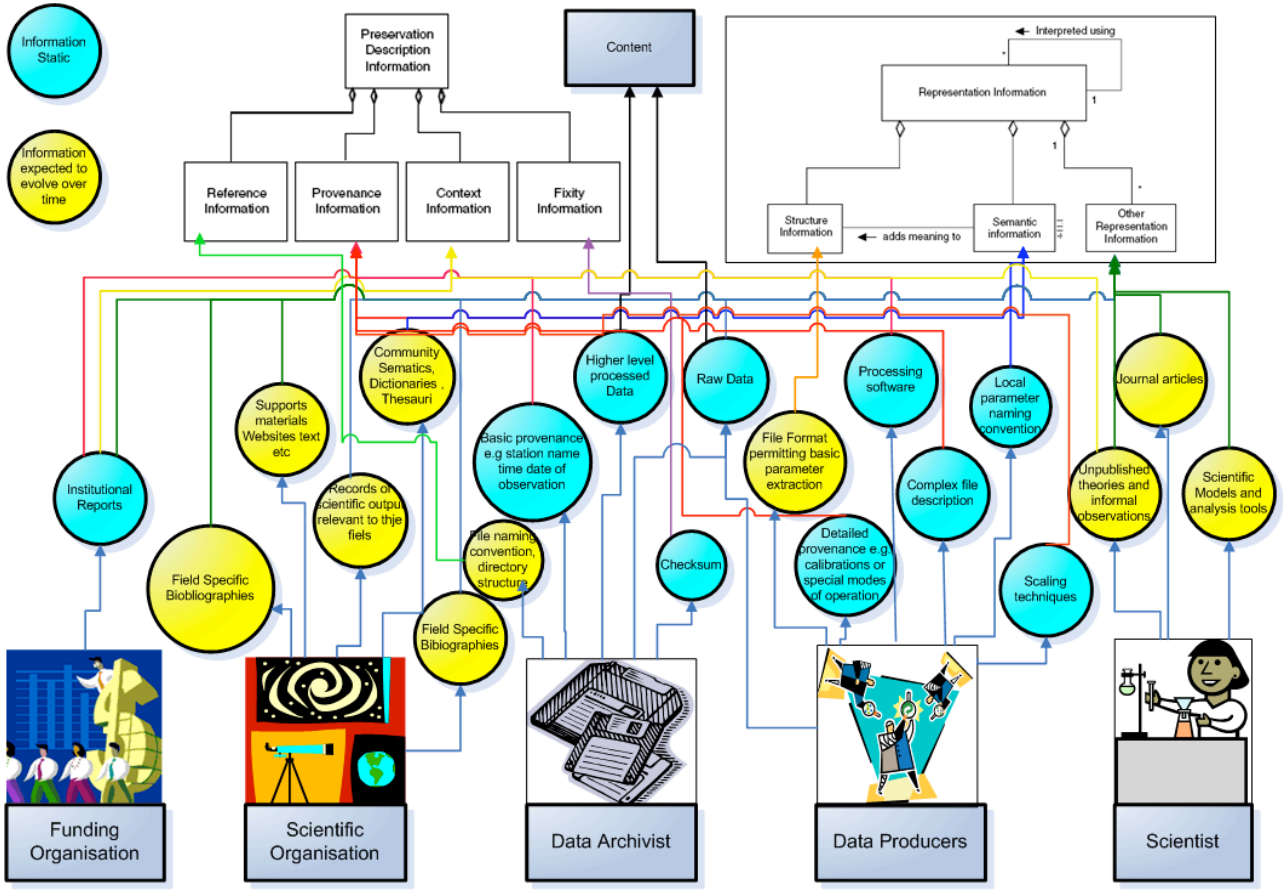


Fig. 2.6 Example OAIS information flow diagram DCC SCARP project

2.6 Preservation strategies

The Information flow diagram should now identify where preservation strategies need to developed, eg in the following areas.

In response to a supply impediment.

Where there is an impediment to the supply, a strategy must be developed in order to either overcome the impediment, either immediately, for example purchasing a special licence for software, or at some later time, for example an institution could develop a simplified open source version of the software which contains the key functionality. The alternative is to develop a mechanism that effectively references the external information object in tandem with a mechanism for monitoring the situation (known as "preservation orchestration").

In response to an identified information preservation risk

Information objects must be inspected on a case by case for their individual preservation risk based on dependencies they have which will be affected by the passage of time. Different strategies which effectively obviate these risks must then be developed.

As a secondary response to a preservation strategy

Where a dependency between information objects has been identified, secondary preservation strategies may need to be developed for related objects.

Multiple strategies can be developed for each instance in these areas. This results in a number of preservation plans being formed.

A preservation plan consists of a unique

- Set of information objects
- Set of supply relationships
- Set of preservation strategies.

The plan allows you to carry out a series of clear actions in order to create an AIP. This allows you to take a number of plans to the cost/benefit stage.

2.7 Cost/Benefit Analysis

Plan options can then be assessed according to

- Costs to archive directly as well as the resources, knowledge and time of archive staff
- Benefits to future users which facilitate re-use of data
- Risks – what are the risks inherent the preservation strategies and are they acceptable to the archive.

Once this analysis is complete the optimal plan can selected and progressed to preservation action see DCC Lifecycle model [21] below. If no plans are deemed suitable then the process must begin again with an adjustment to the preservation objective and/or the Designated Community to be served.

.

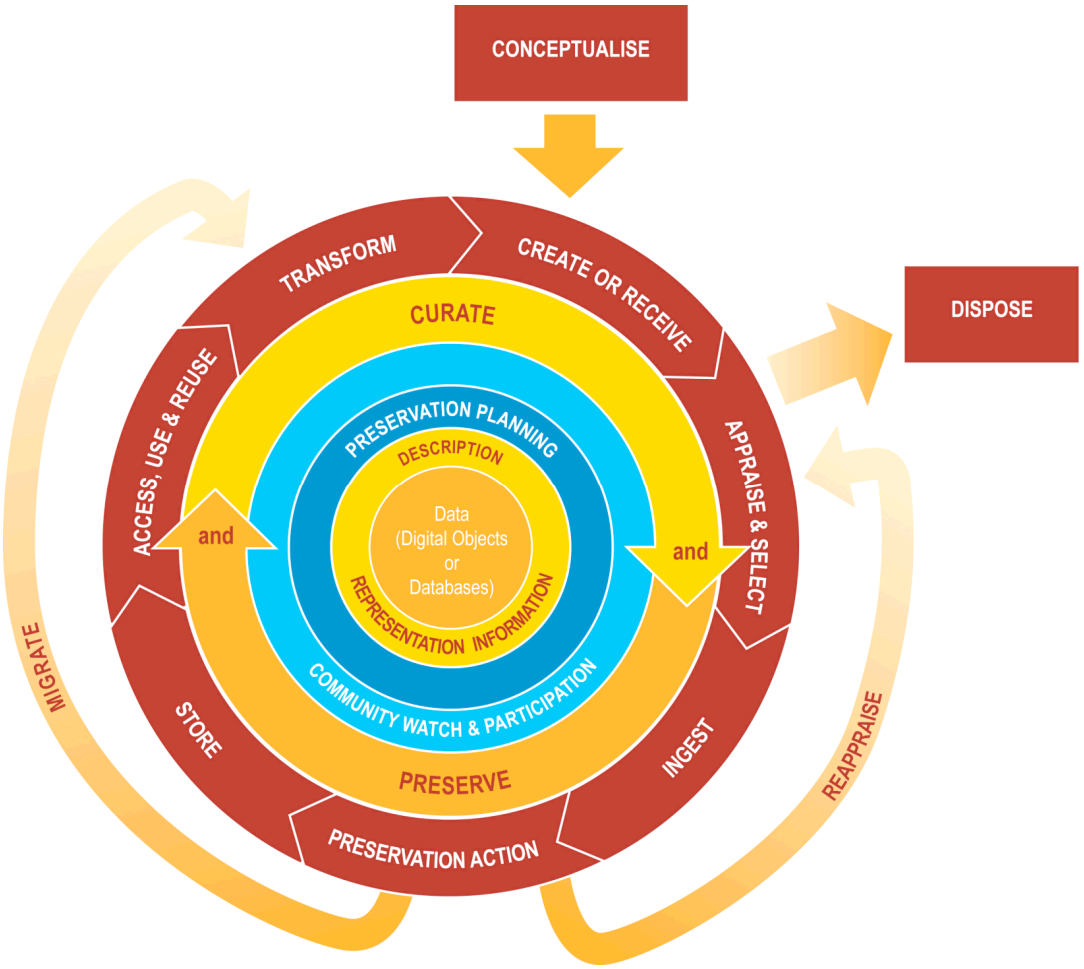


Fig. 2.7 DCC Curation lifecycle model (Higgins, 2008)

3. Analysis applied to the MST data set

MST archive evolution and management

We discussed in the introduction how the MST radar data set was extremely well documented and tightly managed. The result of restricted access and end users being required to report back on how they have used the data is that a record of data use has been created which tracks the evolution of use over time. In addition the Archivist carries out the following key functions

- He is also the project scientist involved in production of the data
- He is a field expert and practising scientist in close contact with relevant scientific organisations, publishing at and attending conferences.
- He additionally provides support, runs and keeps records of user group meetings
- He provides reporting to the funding bodies.

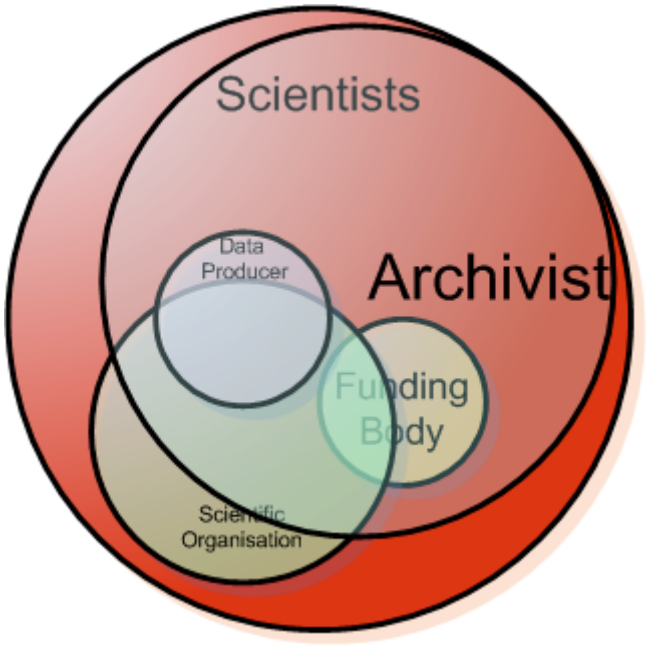


Fig.3.1 Knowledge Overlap MST data stakeholders DCC Scarp project 2008

We have graphically represented the archivist's relationship to other stakeholders to illustrate the overlap of knowledge stakeholders. We see the impact this has on the scope of the preservations objectives we set next in the case study.

MST Preservation Objective

The analysis may present one with a natural easily defined preservation objective or alternatively there may be a greater number of options which overlap and are more difficult to define. For the MST the lowest "buy in" preservation would be that specified in Objective 1 which would permit the simple extraction of parameters from the data.

MST Preservation Objective 1

A user from a future designated user community should be able to extract the following information from the data for a given altitude and time:

- Wind Speed and Direction
- Vertical Velocity
- Radar Return Signal Power
- Tropopause Sharpness
- Tropopause Altitude
- Vertical Wind Sheer
- Beam Broadening Corrected Spectral Width
- Aspect Sensitivity
- Secondary Radial Chain
- Horizontal Wind Complementary Beam Variability
- Horizontal Wind Speed Thetas Compensation Factor.

The data user should also be able to correctly interpret the scientific parameter definitions.

However due to the Archivist's understanding of how data are used by scientists it is possible to carry out a richer level of preservation described in objective 2 which would greatly enhance the value of the long term data archive. The extra effort required would not be prohibitive due to good relationships between the Archivist, data producer, users and the MST International Workshop. Capturing additional information would facilitate future re-use of the data and would give a greater return on investment for effort expended.

MST Preservation Objective 2

A user from a future designated user community should be able to extract the following information from the data for a given altitude and time:

- Wind Speed and Direction
- Vertical Velocity
- Radar Return Signal Power
- Tropopause Sharpness
- Tropopause Altitude
- Vertical Wind Sheer
- Beam Broadening Corrected Spectral Width
- Aspect Sensitivity
- Secondary Radial Chain
- Horizontal Wind Complementary Beam Variability
- Horizontal Wind Speed Thetas Compensation Factor

The data user should also be able to correctly interpret the scientific parameter definitions. This objective is extended by saying we wish future users to be able access, read and interpret

- Scientific output resulting from use of the MST data set
- The MST international workshop conference proceedings
- The MST user group meeting minutes.

In the methodology we stated that preservation objectives should be

- Specific: well defined and clear to anyone with a basic knowledge of the domain
- Actionable: the objective should be currently achievable. It is important to note the information ultimately to be extracted by a future user cannot be predicted and therefore we should not attempt to "predict the future"
- Measurable: it is critical to be able to know when the objective has been attained in order to assess if any preservation strategy developed is adequate.

As a result we specified the information we wanted the user to be able to access and utilise. It may have been tempting to set objective as being a desire for future users to be able to

identify common atmospheric phenomena with the same degree of skill as current users. We should however not do this, as it not clearly actionable and we would not know with any degree of certainty if this had been achieved.

MST Designated user Community

The Designated Community is defined in OAIS as **"**An identified group of potential Consumers who should be able to understand a particular set of information. In the case of the MST data the designated user community which the archive wishes to serve is that of the UK atmospheric science research community. It believes with reasonable confidence the community will still

- Possess the basic knowledge of a UK physics graduate
- be able to read English
- be numerate and able to acquire necessary skill to meaningfully manipulate, analyse and create models for data
- have sufficient technical skills within the community to write programs or scripts to extract parameters from data files given adequate structural description
- be able to comprehend current journal literature on atmospheric science.

The Designated Community will possess a skills and knowledge base which allows them to successfully interact with a set of information which has been stored within an AIP in order to extract required knowledge or recreate the required performance or behaviour. As a result the Designated Community determines the information which must essentially be contained by the AIP but also the form this information takes and the optimal preservation strategies to be employed. We will discuss the implications of this Designated Community definition when multiple strategies present themselves in section 3.5.

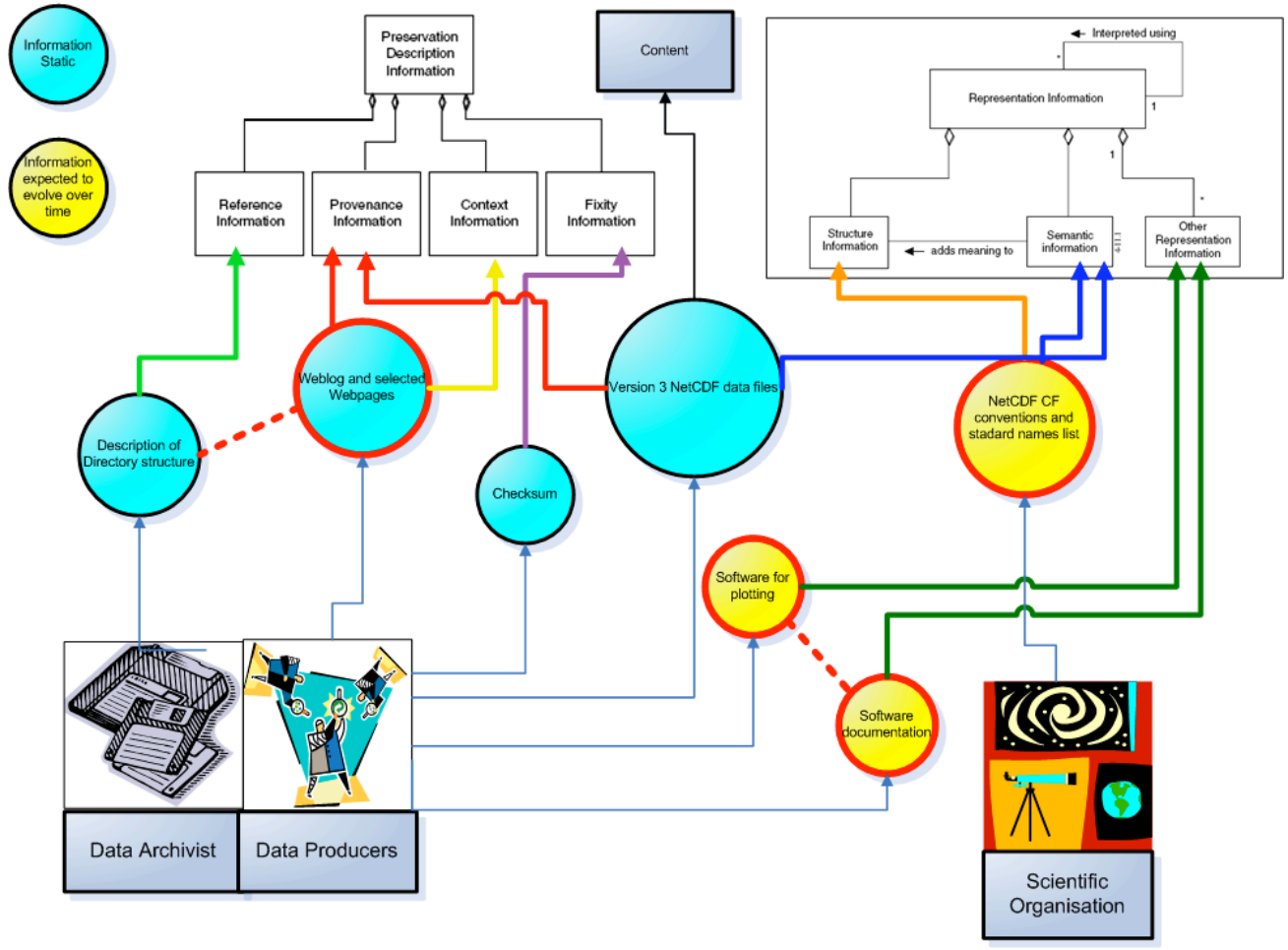


Fig. 3.2 Preservation Information flow for objective one DCC Scarp project 2009

3.1 Content – MST Version 3 NetCDF data files:

NetCDF (network Common Data Form) [22] is an interface for array-orientated data access and a library that provides an implementation of that interface. NetCDF is used extensively in the atmospheric and oceanic science communities. It is a preferred file format of the British Atmospheric data centre who currently provide access to the data. The NetCDF software was developed at the Unidata Program Center in Boulder Colorado USA [23] http://www.unidata.ucar.edu/. NetCDF facilitates preservation for the following reasons

- NetCDF is a portable, self-describing binary data format so is ideal for capture of provenance, descriptive and semantic information.
- NetCDF is network-transparent, meaning that it can be accessed by computers that store integers, characters and floating-point numbers in different ways. This provides some protection against technology obsolescence.
- NetCDF datasets can be read and written in a number of languages, these include C, C++, FORTRAN, IDL, Python, Perl, and Java. The spread of languages capable of reading these datasets ensures greater longevity of access because as one language becomes obsolete the community can move to another.
- The different language implementations are freely available from the UNIDATA Center, and NetCDF is completely and methodically documented in UNIDATA's NetCDF User's Guide making capture of necessary representation information a relatively easy low cost option.
- Several groups have defined conventions for NetCDF files, to enable the exchange of data. BADC has adopted the Climate and Forecasting (CF) conventions for NetCDF data and have created standard names.

CF conventions are guidelines and recommendations as to where to put information within a NetCDF file, and they provide advice as to what type of information you might want to include. CF conventions allow the creator of the dataset to include information representation and preservation description information in a structured way. Global attributes describe the general properties and origins of the dataset capturing vital provenance and descriptive information, while local attributes are used to characterise the recorded variables thereby capturing the all necessary semantics.

3.2 Checksum

The BADC currently runs a checksum program on its archived data every 80 days. The data checksum should be checked before ingest at which point it would be replaced by a new checksum as part of the packaging solution. This could be provided by the DCC/CASPAR developed packaging tool [21] http://developers.casparpreserves.eu:8080/hudson/job/CASPAR-PACK/.

3.3 Weblog and Selected WebPages

Much additional valuable provenance information has also been recorded in the MST radar support website. Selected pages or the entire site could be archived as Preservation Description Information.

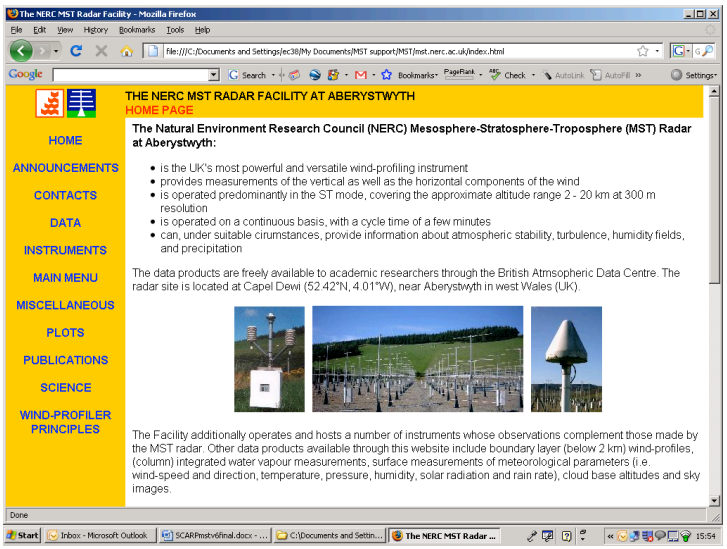


Fig 3.3 MST Website

The MST website is currently located at http://mst.nerc.ac.uk/ . Due to the sites' simple structure, which consists of a set of static pages and common file types it would be a relatively simple operation to run a web archiving tool such as HTtrack[25] http://www.httrack.com/ to copy the website and add additional repinfo on HTML, PDF, postscript, MS Word and JPEG from the DCC Registry Repository of Representation Information RRORI. HTtrack is only one of a range of web-archiving tools which are freely available and require minimal skill to operate. However it is worth noting that it is only by virtue of the technical simplicity of the site that it is so relatively easy to archive and preserve.

The instrument performance weblog also provides essential provenance as it details any problems with the instrument (see sample entry below)

[Data]

Number	157
Start	2006-08-05 17:20
End	2006-08-05 21:10
Added	2006-08-07 08:54
Author	DAH

Instrument: MST Radar. Interference was experienced, primarily from 17:20 - 21:10 UT and apparently confine vertical beam observations. It was most obvious at the lower range gates of the ST mode. It appears to have resulted in missing data for the v2 processing, rather than contamination. A more minor incident occurred aroun 11:10 UT.

The log is currently accessed at http://mst.nerc.ac.uk/cgi-bin/mstlog_public/mst_event_search.py . The information is stored in a Postgres database but due to the simple structure of the data base table and relatively small volume of information, dumping it into a CSV file is a practical low cost option that would provide greater longevity.

3.4 Description of directory structure and BADC file naming conventions

The current directory structure is logical and well thought out. This should be maintained in the AIP package. Details of archiving conventions are recorded in the MST website http://mst.nerc.ac.uk/archiving_conventions.html which will need to be altered by the removal of the BADC from the top of the directory hierarchy structure to avoid confusion.

/badc/dataset-name/data/data-type-name/YYYY/MM/DD/

3.5 MST access and plotting Software with accompanying documentation

The BADC considers the long term archiving of software to be an impractical option principally due to the complex dependencies of software. It takes the view that it expects much current software to be superseded by newer software which will be capable of recreating and enhancing much of the existing analysis and access functionality.

MST data plotting software

An example of data set specific plotting and analysis programs for the MST would be the MST GNU plot software. This software plots Cartesian product of wind profiles from NetCDF data files and was responsible for figures 1.5 – 1.8 above . This software was developed by the project scientist due to specialised visualization requirements where finer definition of colour and font was needed than that provided by generic tools.

Preservation risks are due to the following user skill requirements and technical dependencies.

- UNIX [26]http://www.unix.org/ or Linux distribution
- The user must be able to install python[27] http://www.python.org/ with python-dev module installed with numpy array package and pycdf
- GNU plot to be installed [28] http://www.gnuplot.info/docs/gnuplot.html and a user must be able to set environmental variables
- The ability to run required python scripts through a UNIX command line
- GNU plot template file to format plot output.

A number of preservation strategies present themselves,

Emulation strategy

One solution is preserving the software through emulation, for example Dioscuri [29] http://dioscuri.sourceforge.net/faq.html. Current work with the PLANETS project [30]http://www.planets-project.eu/news/?id=1190708180 will make Dioscuri capable of running operating systems such as Linux Ubuntu which should satisfy platform dependencies. With the capture of specified software packages/libraries and the provision of all necessary user instructions this would become a viable strategy.

Conversion strategy

It is additionally possible to convert NetCDF files to another compatible format such as NASA AMES[31] http://badc.nerc.ac.uk/help/formats/NASA-Ames/ . We were able to achieve this conversion using the community developed software Nappy [32] http://home.badc.rl.ac.uk/astephens/software/nappy/ , CDAT[33] http://www2-pcmdi.llnl.gov/cdat and Python. This is a compatible self describing ASCII format, so the information should still be accessible and easily understood as long as ASCII encoded text can still be read. There would be however reluctance to do this as NASA AMES files are not as easily manipulated making it more cumbersome to analyse data in the desired manner.

Preservation by addition of Representation information strategy

An alternate strategy is to gather the following documentation relating to the NetCDF file format which contains adequate information for future users to extract the required parameters from the NetCDF file.

Currently this information can be found in the BADC support pages on NetCDF http://badc.nerc.ac.uk/help/formats/NetCDF/ [34] which can be archived using the HTtrack tool or adequately referenced. These pages suggest some useful generic software a future user may wish to utilize.

If these pages or no longer available or the software is unusable a user can consult documents from the NetCDF documentation and libraries from Unidata http://www.unidata.ucar.edu/software/NetCDF/docs/ [35]. This means that if future user community still have skills in FORTRAN, C, C++ or Java they will be able to easily write software to access the required parameters.

3.6 CF Standard names list

The NetCDF files contain semantic descriptions of the parameters as part of the file header. The parameter names and descriptions will be susceptible to semantic drift over time. By inclusion in the standard names list http://cf-pcmdi.llnl.gov/documents/cf-standard-names/about the quality of the description will have been scrutinized by the current user community. Some preservation orchestration to monitor this list over to time in order to guard against semantic drift would be required.

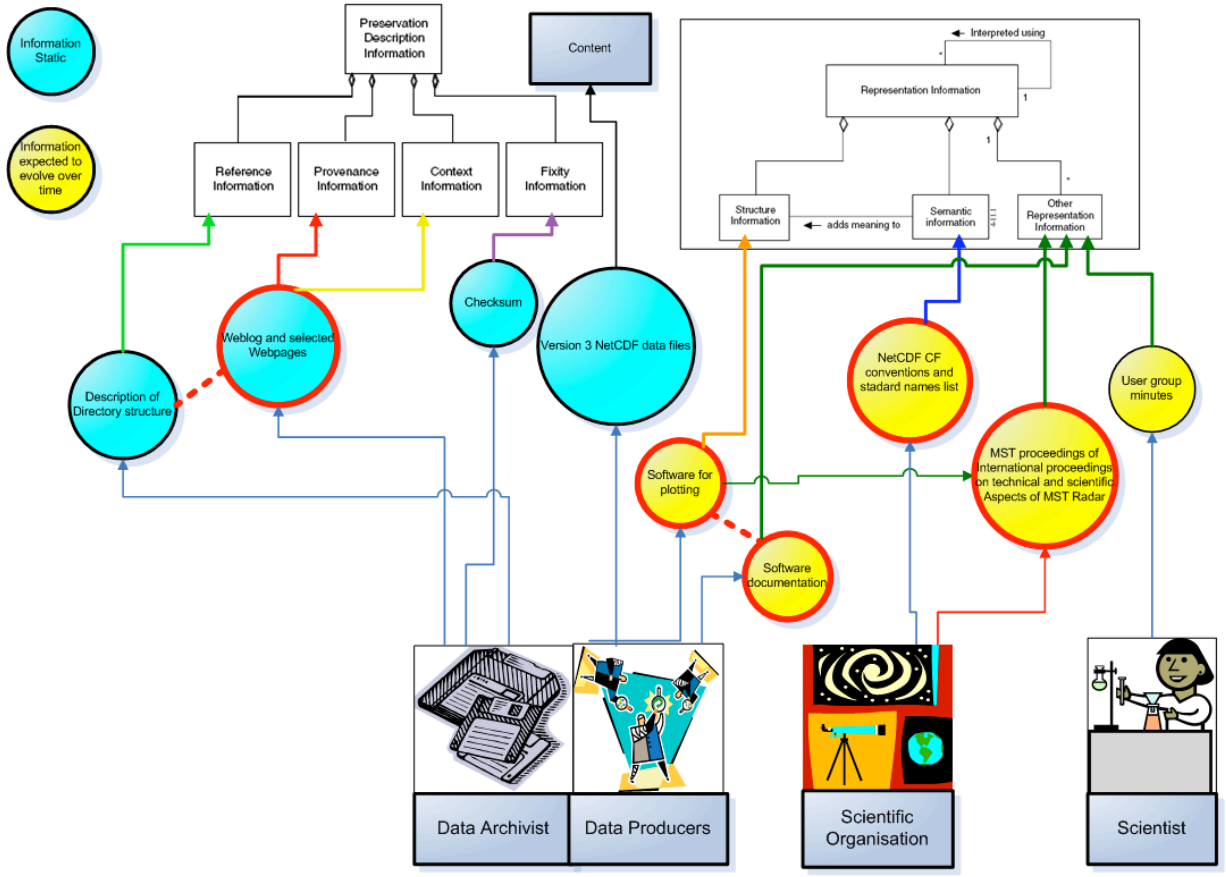


Fig. 3.4 Preservation Information flow for objective two ©DCC SCARP project 2009

Satisfying preservation objective two requires the inclusion of three additional sets of information which permit "the identification of common atmospheric phenomena which have been previously established by MST data users and noted in peer reviewed literature or the MST conference proceedings". The information entities from preservation objective one and their associated preservation strategies remain the same.

3.7 User Group minutes

The project scientist has again been quite diligent in keeping minutes of the user group meetings which are run for data-using scientists several times a year. As result this information is easily captured. It currently resides in the NCAS [36] ceda repository which provides easy access to current data users however there are no guarantees that this repository will persist in the longer term so a simple reference in the form of URL would not be considered to be sufficient to guarantee permanent access to this material. This leaves two strategies open to the archive. The first involves taking a copy of this material and including it physically within the AIP. The second involves orchestration where the ceda repository would be required to alert the custodians of the MST data to the demise of the repository or migration of this material, so it may be obtained for direct inclusion in the AIP.

3.8 Record of scientific output

The website additionally contains a bibliographic record of publications resulting from use of the data. This record contains good quality citations but there would be concerns regarding permanent access to some of these materials, consider the two examples below

W. Jones and S. P. Kingsley. MST radar observations of meteors. In *Proceedings of the Wagstaff (USA) Conference on Astroids, Comets and Meteors*. Lunar and Planetary Institute (NASA Houston), July 1991

S. P. Kingsley. Radio-astronomical methods of measuring the MST radar antenna. Technical report to MST radar user community, 1989.

Neither of these two items are current held by either The British Library http://www.bl.uk/ [37] or The Library of Congress http://catalog.loc.gov/ [38] based on searches of their catalogues. Nor do they in exist in the local STFC institutional repository http://epubs.cclrc.ac.uk/ [39]

A preservation strategy to deal with this bibliography would be to create MARC [40]http://www.loc.gov/marc/ [41] http://www.dcc.ac.uk/diffuse/?s=36 records in XML format for items held by the British Library and to obtain copies of the other items from the current community and digitise them in PDF format for direct inclusion within the AIP.

3.9 Proceedings for the International workshop on the technical and scientific aspects of MST radar

The international workshop on MST radar is held about every 2-3 years, and is a major event gathering together experts from all over the world, engaged in research and development of radar techniques to study the mesosphere, stratosphere and troposphere (MST). It was additionally attended by young scientists, research students and also new entrants to the field to facilitate close interactions with the experts on all technical and scientific aspects of MST radar techniques. It is this aspect which makes the proceedings an ideal resource for future users who are new to the field.

Permanent access to these proceedings is again at risk. The MST 10 proceedings are available for download from the internet [42] http://jro.igp.gob.pe/mst10/ and from the British Library. Proceedings 3, 5-10 are also available from the British library, meeting 4 is only available from the Library of Congress and unfortunately the proceedings from meetings 1 and 2 have not been deposited in either institution.

Again a number of strategies present themselves. Copies of proceedings 1, 2 and 4 could be obtained from the still active community, digitised and incorporated into the AIP. The proceedings which are currently held by the British Library can be obtained, digitised and incorporated into the AIP or alternatively the XML MARC record can be obtained and incorporated into the AIP as a reference as there is a high to degree of confidence in the permanence of these holdings.

4. Conclusions and Recommendations

In this Case Study we have applied a preservation analysis methodology which is discipline independent in application but none the less capable of identifying and drawing out discipline specific preservation requirements and issues. In this section we examine the implications of issues raised by the case study both for and outside the immediate needs of the MST data set and atmospheric sciences community. We additionally make recommendations to the DCC in order to support curation and preservation outside the discipline, in order to serve the wider community

MST data stakeholders share a mutual characterisation of stakeholders with other scientific disciplines. It is significant for this dataset that an individual holds three stakeholder positions. The comparatively small size of the user community, and the tight management and close interaction of stakeholders are all factors which facilitate information capture and a more comprehensive understanding of the field, data re-use and knowledge the data is capable of imparting. These factors along with conscientious documentation and good data curation/management practice have had the effect of extending the preservation objective which may be achieved. It is however difficult to attach value to this extension. More research into building quality business cases for scientific data would be beneficial for archives presented with a number of potential preservation objectives.

Different knowledge or information may be extracted from a single data set. The second extended preservation objective in this case study results from the fact that the data is a real observation of a dynamic system (the atmosphere). Understanding the established processes within such a system will naturally inform the re-use of the data. This makes a preservation objective which involves the capturing this level of understanding highly desirable. This would not be discipline specific as we would expect other scientific datasets to chow similar characteristics e.g. those in oceanographic, ecological or economics domains where re-use of data builds on previous work carried out be fellow scientists within the community.

The designated community will possess a combination of both discipline specific and more general skills and knowledge base. This combination will however produce a unique profile against which the adequacy of information contained within an AIP can be assessed. In the case of the MST radar user community there is the assertion that the community will maintain sufficient numeric, analytical and technical skills to extract and manipulate parameters if the structure is fully documented. The ability to make such assertions relies heavily upon discipline specific knowledge and awareness of the user community. This has implications for any OAIS based audit process as this introduces an element of trust in the assertions by made discipline. An auditor may request clarification but could find it difficult to appraise assertions that are heavily reliant upon discipline specific knowledge.

While each element of the OAIS information model (provenance, structure etc) will be universally required, the unique set of constituent information objects will be dependent on the dataset, the preservation objective and the defined designated user community. Adoption of similar data management practices and technologies within a discipline will mean there is a degree of re-use of Representation Information across that discipline. We would for example expect structural information for NetCDF files to be widely reused in AIPs across the atmospheric science domain. The

deposit of such information in an accessible registry repository such as RRORI would therefore be highly beneficial for the wider community.

Data curation practices developed within the community such as the adoption of self describing file formats, documentation of the MST site at Aberystwyth and documenting activities of the user community facilitate easier preservation now. However this is has not been adequate to completely ensure preservation for this data set; further action needs to be taken. Strategies still need to be developed and appraised for objects with preservation risks on a case by case basis. Some strategies will emerge from the community for example

- Adoption of community based semantic quality control using CF standard names list in conjunction with the CASPAR preservation orchestration manager
- Conversion of NetCDF to a NASA AMES format using community developed software CDAT, NAppy and BADC validation tools
- Addition of community created representation information on NetCDF from the BADC and UNICAR.

Other strategies have been developed through digital curation and preservation projects for example

- Emulation of Linux Ubuntu by PLANETS/DIOSCURI to preserve software
- Capture of provenance and community knowledge using web archiving tools
- Addition of representation information from the DCC registry repository RRORI
- Use of the DCC/CASPAR Packaging Tool to physically create the API.

Due to the fact that valid preservation strategies come from a variety of sources that are not always obvious, a service which informs archives of the range of suitable option for common information objects would be highly desirable. Quality assurance and testing of these solutions would also be desirable.

Recommendations

Recommendation for consideration by the Archive

According to the NERC Data Policy Handbook "A review mechanism must exist to reconsider periodically the cost benefits of continuing to maintain the data. The intention to destroy or put at risk data should be publicised in advance, allowing time for a response by interested parties." Within the BADC efforts to preserve data long term are encouraged but are not mandated. Given this it is our recommendation that the archive create a preservation plan based on a cost benefit and risk assessment of the available strategies. Publish this along with an assessment of the preservation objective and the Designated Community for public scrutiny and comment. Review this plan periodically altering it in response to environmental changes and improvements in preservation techniques. Carry out necessary preservation action and create a "logical Archival Information package" during this current period of quality management and active use while the resources and information are still obtainable. This should allow the data to be retired from active management or transferred to another organisation with greatly reduced preservation risk.

Recommendation for consideration by the DCC and wider community

It was felt that there is a need to support preservation analysis and planning at the data set level and establish a process which is comprehensive and aware of all elements required for the re-use of data in the long term. We also identified areas where archives may benefit from external support in order carry out appropriate analysis, strategy selection and preservation action.

Recommendation ~1 Wider application, trialling and further development of the preservation analysis methodology outlined here would be desirable to test its validity in a broader range of disciplines and organisational settings. In addition the production of training materials and support for archivists who wish to adopt our approach for data preservation would be of benefit.

Recommendation ~ 2 We view the preservation analysis methodology as complimentary to repository planning, audit and certification activities. Further investigation is needed into how the results of preservation analysis could be fed into audit and risk analysis assessments such as Drambora. Integration of preservation analysis with other digital preservation practices is necessary to provide archives caring for scientific data sets with the full arsenal of tools and techniques necessary to rise to the challenge of digital preservation.

Recommendation ~3 Archives can find it difficult to articulate and specify reasons for the preservation of data. We recommend that the DCC develops further guidance on setting preservation objectives and establishing valid business cases for the preservation of scientific data.

Recommendation ~4 Archives need to establish the skill and knowledge base they should monitor in their "designated community", in order to ensure data re-use. The DCC should investigate this area further, and provide guidance and assessment tools to facilitate the meaningful definition and monitoring of such a designated community.

Recommendation ~ 5 Persistent access to grey literature that supports data re-use is an important issue. Advice on approaches for the deposit and citation of such material would be a valuable service for archives.

Recommendation ~ 5a Similarly, the data curation community would benefit from an notification service for repositories that are in danger of closing or whose content is being migrated to another repository to ensure persistent access to required content.

Recommendation ~6 DCC could offer an advisory service which recommends or provides information on preservation strategies available to archives. It could additionally provide quality assurance and testing for representation information deposited in the DCC Registry/Repository of Representation Information (RRORI).

Recommendation ~7 The DCC or another identified organisation could provide an archiving service and/or assistance for web based collections of representation information and preservation description information.

Recommendation ~8 The MST data has benefitted from many good data management practices recommended through the British Atmospheric Data Centre. Other data sets from outside the atmospheric sciences could benefit from similar approaches. Self-describing, well documented data formats such as NetCDF, semantic control through CF standard name conventions, and software development initiatives are just some examples of practices which could be transferred outside the discipline. The DCC should play an instrumental role in transferring good practices between disciplines.

Acknowledgements

We acknowledge the assistance and support from David Hooper , Sam Pepler, Alan Harwood, David Giaretta, Simon Lambert, Brain Matthews , Matthew Dunckley, Stephen Rankin and Brian McIlwrath.

References

[1] Met Office http://www.metoffice.gov.uk/
[2] Natural Environment Research Council http://www.nerc.ac.uk/
[3] Rutherford Appleton Laboratory http://www.scitech.ac.uk/About/Find/RAL/Introduction.aspx
[4]British Atmospheric Data Centre http://badc.nerc.ac.uk/home/index.html
[5]National centre for Atmospheric Science http://www.ncas.ac.uk/
[6]NERC data policy handbook http://badc.nerc.ac.uk/data/NERC_Handbookv2.2.pdf
[7]MST Website http://mst.nerc.ac.uk/
[8]CASPAR Questionnaire http://www.casparpreserves.eu/Members/cclrc/ReferenceDocuments/caspar-test-case-questionnaire/at_download/file
[9] - Consultative Committee for Space Data Systems. ReferenceModel for an Open Archival Information System (OAIS). CCSDS 650.0-B-1. Jan 2002 http://public.ccsds.org/publications/archive/650x0b1.pdf
[10] Ionosonde Case Study http://www.casparpreserves.eu/other-caspar-products/other-caspar-products/ionosonde-case-study.pdf
[11] A. J. McDonald, K. P. Monahan KP, D. A. Hooper, and C. T. Gaffard. VHF signal power suppression in stratiform and convective precipitation. Ann. Geophys., 24:23-35, 2006.
[12] - R. Bose, R. Mann and D. Prina-Ricotti, 2006. AstroDAS: Sharing Assertions across Astronomy Catalogues through Distributed Annotation (PDF).
[13] - CLADDIER project http://www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories/project_claddier.aspx

[14] - OJIMs Project
http://www.jisc.ac.uk/whatwedo/programmes/programme_rep_pres/repositories_sue/ojims.aspx
[15] EISCAT http://www.eiscat.rl.ac.uk/
[16] World Data Centre at the Rutherford Appleton Laboratory
http://www.wdc.rl.ac.uk/wdcc1/wdc_menu.htmlb
[17] NetCDF Climate and Forecast Metadata convention http://www.cfconventions.org/
[18] - Ionospheric Parameter Codes http://www.ukssdc.ac.uk/wdcc1/ionosondes/ursi_codes.html
[19] - W. R. Piggott and K. Rawer . URSI handbook of ionogram interpretation and reduction. UAG 1972 http://www.ips.gov.au/IPSHosted/INAG/uag.htm
[20] XFDU http://sindbad.gsfc.nasa.gov/xfdu/index.html
[21] DCC Lifecycle Model Higgins2008 http://www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf
[22] NetCDF http://www.unidata.ucar.edu/software/NetCDF
[23] *UNICAR* http://www.unidata.ucar.edu
[24] DCC/CASPAR packaging tool
http://developers.casparpreserves.eu:8080/hudson/job/CASPAR-PACK/
[25]HTT track website copier http://www.httrack.com/
[26] The Unix open group http://www.unix.org/
[27] Python Programming Language official website http://www.python.org/
[28] GNUplot website http://www.gnuplot.info/
[29] Dioscuri on Sourceforge http://dioscuri.sourceforge.net/
[31] Ubuntu emulation by Dioscuri http://www.planets-project.eu/news/?id=1190708180
[32] NASA Ames http://badc.nerc.ac.uk/help/formats/NASA-Ames/
[33]Nappy http://home.badc.rl.ac.uk/astephens/software/nappy/
[34]CDAT http://www2-pcmdi.llnl.gov/cdat
[35] BADC helppages on NetCDF http://badc.nerc.ac.uk/help/formats/NetCDF/
[36] NetCDF document library http://www.unidata.ucar.edu/software/NetCDF/docs/
[37] NCAS CEDA repository http://cedadocs.badc.rl.ac.uk/
[38] British Library http://www.bl.uk/
[40]Library of Congress http://catalog.loc.gov/
[41]STFC institutional repository ePubs http://www.scitech.ac.uk/Publications/lib/ePubs.aspx
[42] Library of congress MARC standards http://www.loc.gov/marc/
[43]DCC diffuse standards framework MARC 21 http://www.dcc.ac.uk/diffuse/?s=36
[44]MST 10 workshop proceeding http://jro.igp.gob.pe/mst10/